DBARF: Deep Bundle-Adjusting Generalizable Neural Radiance Fields - Supplementary Materials

Yu Chen Gim Hee Lee
Department of Computer Science, National University of Singapore
{chenyu, gimhee.lee}@nus.edu.sg

1. Additional Implementation Details

Feature extraction network architecture. We use ResNet34 [3] as our feature extraction network backbone. The detailed network architecture is given in Fig. 1. Note that after up-convolution, the shape of resulted feature map may not be the same as the shape of the feature map to be concatenated. In this case, we use bilinear interpolation to rescale them to the same shape.

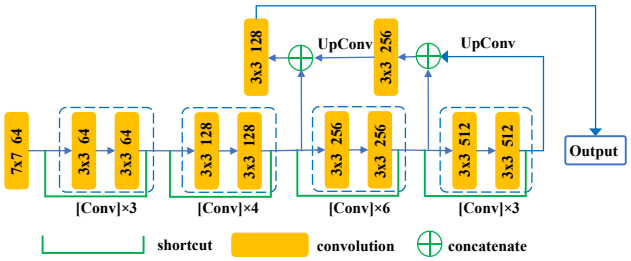

Figure 1. Feature extraction network architecture. The max-pooling layer after the first convolution layer is omitted. Each residual block is a standard block in ResNet [3]. "Conv" denotes convolution, "UpConv" denotes up-convolution.

Depth sampling in GeNeRF. For both IBRNet and our method, we sample 64 inverse-depth samples uniformly. For BARF [4] and GARF [1], the number of inverse depth samples is 128.

2. Results on LLFF Dataset

We present more qualitative results of the LLFF forward-facing dataset [5] in Fig. 2 and more geometry visualization results in Fig. 3.

3. Results on IBRNet's Self-Collected Dataset

To further support the effectiveness of our method, we evaluate our method on more IBRNet's self-collected dataset [6]. The qualitative results of the rendered image and the corresponding geometry after finetuning are respectively given in Fig. 5 and Fig. 6. The quantitative results are given in Table 1. We notice there is a drop in PSNR for 'usps van'. This is because a large portion of the images in 'usps van' contains the shadow of a moving human (see Fig. 4), and our method cannot handle moving objects.

Scenes	PSNR ↑		SSIM ↑		LPIPS ↓	
	Ours	Ours_{ft}	Ours	Ours_{ft}	Ours	Ours_{ft}
red corvette	18.04	19.46	0.728	0.785	0.345	0.238
usps van	16.96	16.68	0.761	0.766	0.258	0.208
path lamp & leaves	19.10	20.34	0.403	0.56	0.497	0.275
purple hyacinth	18.19	19.48	0.387	0.506	0.448	0.269
artificial french hydrangea	17.61	19.03	0.531	0.600	0.470	0.292
red fox squish mallow	23.26	24.37	0.668	0.700	0.358	0.270
mexican marigold	20.59	21.48	0.506	0.582	0.419	0.240
stop sign	21.27	22.04	0.738	0.803	0.195	0.099

Table 1. Quantitative results of novel view synthesis on ibrnet self-collected dataset [6]. Ours_{ft} denotes our results after finetuned $60,000$ iterations.

4. Results on ScanNet Datasets

We also show the results evaluated on 7 scenes of the ScanNet dataset [2]. The qualitative results of the rendered image and the corresponding geometry after finetuning are respectively given in Fig. 7 and Fig. 8. The quantitative results are given in Table 2. We can observe that our method outperforms IBRNet by a large margin. The poor performance of IBRNet on this dataset is due to the inaccurate camera poses. However, our method does not rely on pre-computed camera poses and the regressed camera poses are accurate to enable high-quality image rendering.

Scenes	PSNR ↑		SSIM ↑		LPIPS ↓	
	IBRNet [6]	Ours	IBRNet [6]	Ours	IBRNet [6]	Ours
scene0671-00	12.29	26.60	0.518	0.910	0.451	0.113
scene0673-03	11.31	23.56	0.457	0.859	0.615	0.156
scene0675-00	10.55	19.95	0.590	0.875	0.589	0.207
scene0680-00	14.69	31.05	0.709	0.958	0.389	0.056
scene0684-00	18.46	33.61	0.737	0.975	0.296	0.052
scene0675-01	10.33	23.56	0.595	0.899	0.548	0.166
scene0684-01	14.69	33.01	0.678	0.967	0.426	0.056

Table 2. Quantitative results of novel view synthesis on ScanNet dataset [2] after finetuning.

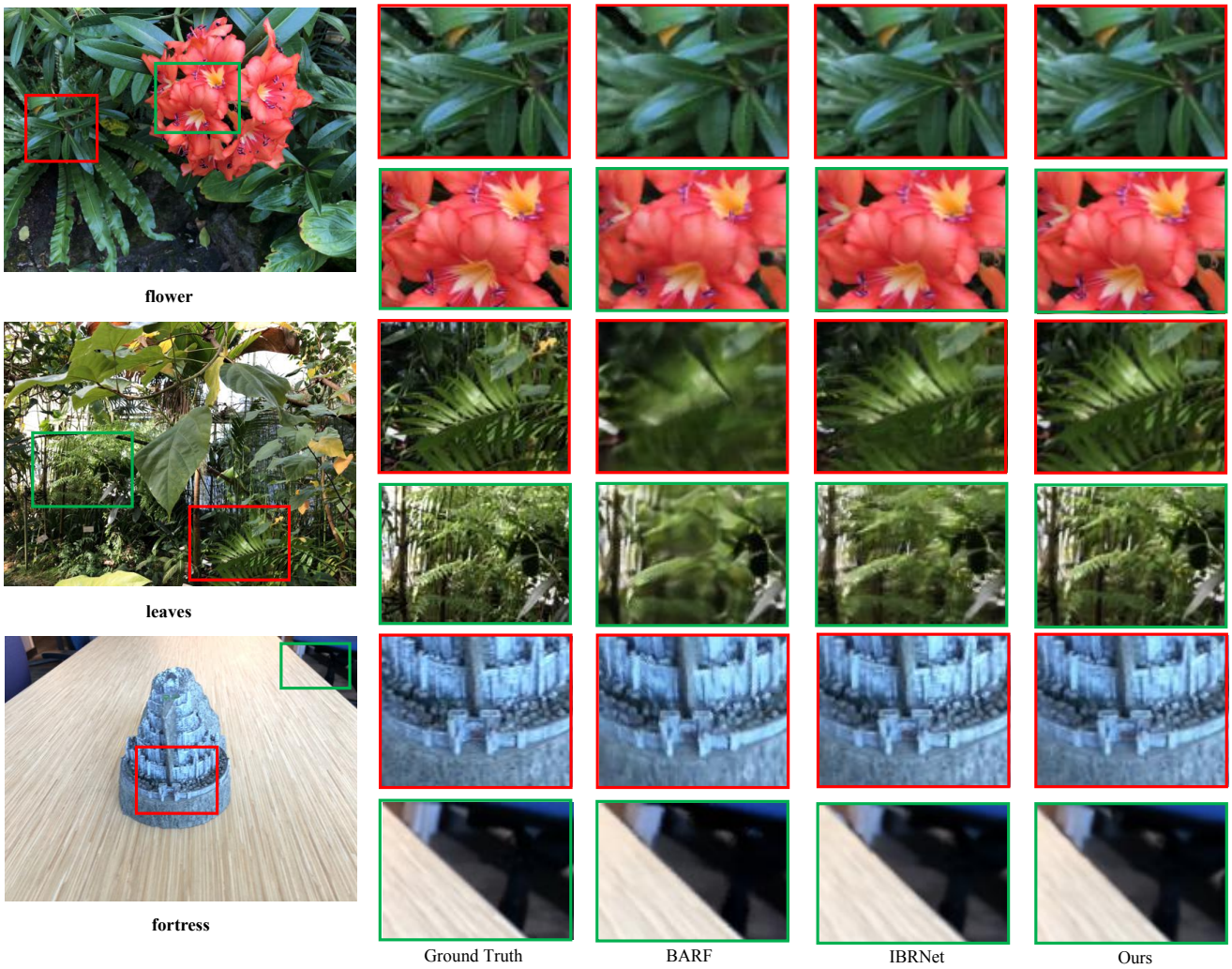

Figure 2. **The qualitative results on LLFF forward-facing dataset [5]**. We show the finetuned results for IBRNet and Ours.

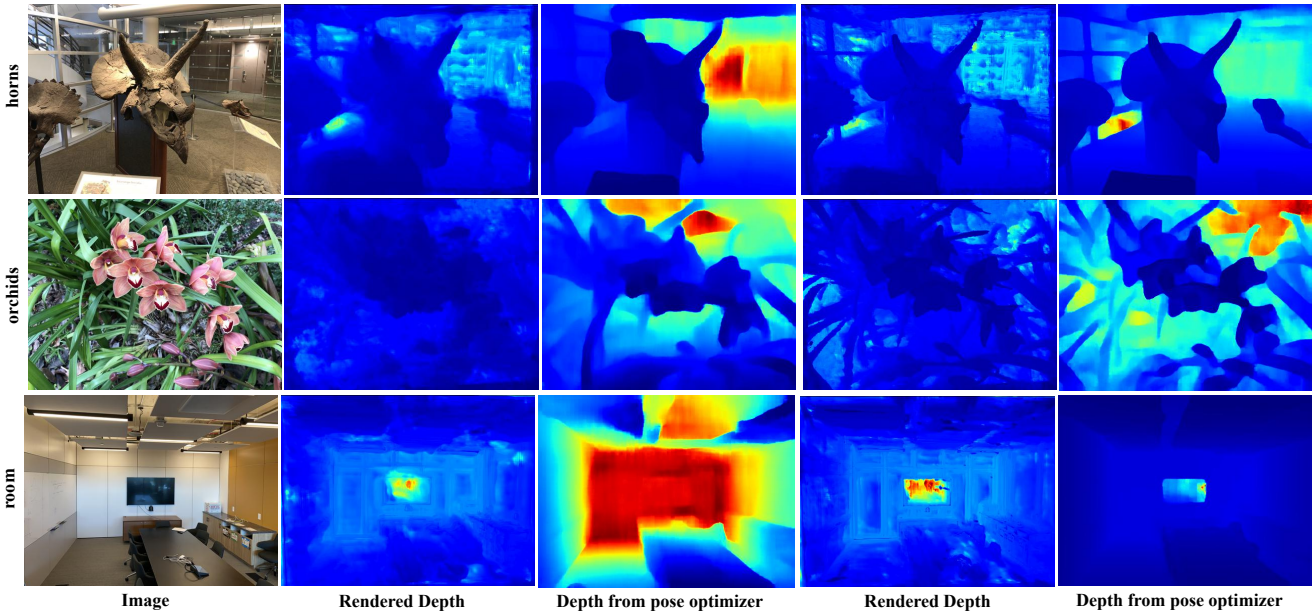

Figure 3. **Depth visualization on LLFF forward-facing dataset [5]**. We show the finetuned results for IBRNet and Ours.

Figure 4. A moving human shadow on the ‘usps van’ scene.

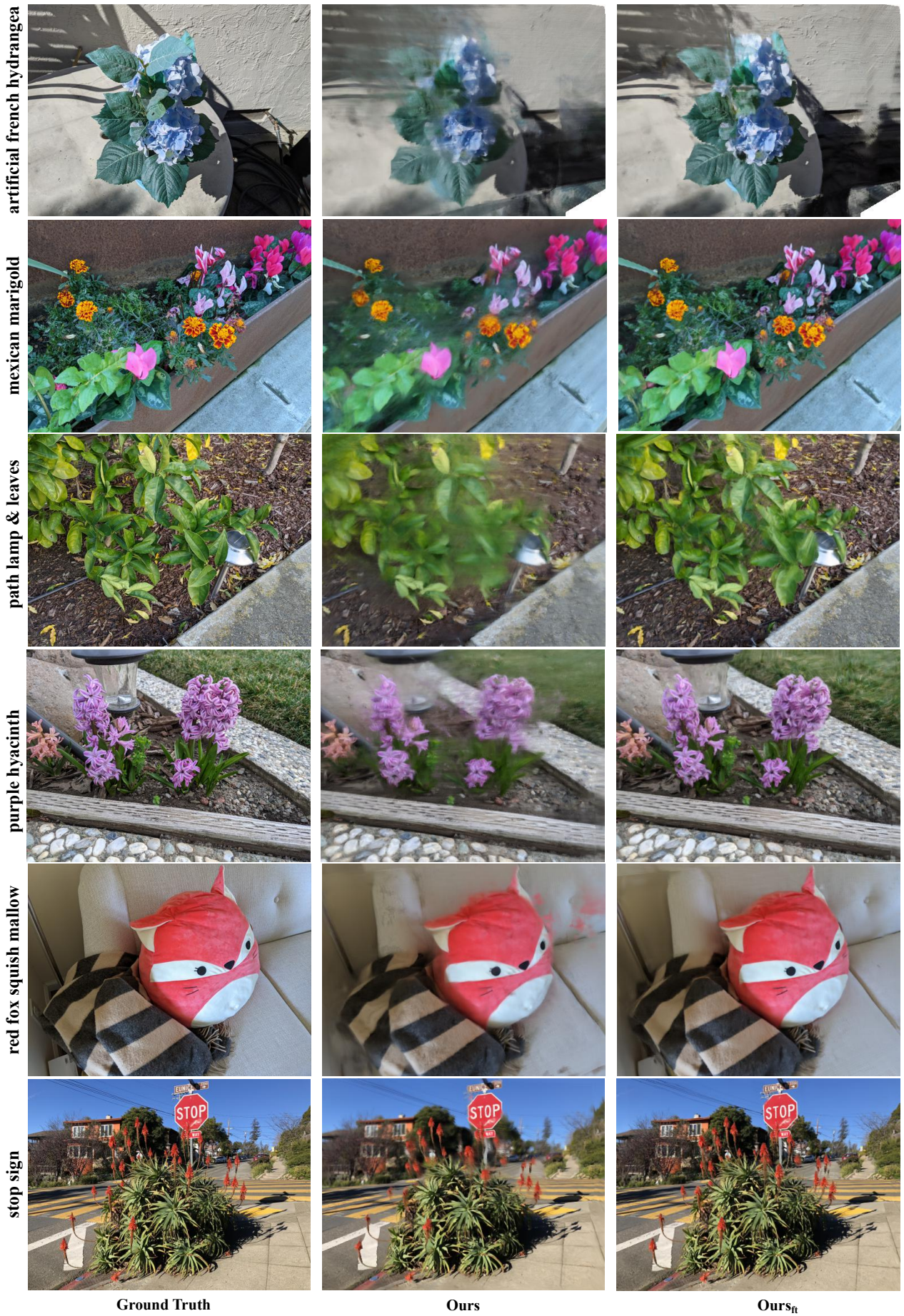

Figure 5. **The qualitative results on ibrnet self-collected dataset [6].**

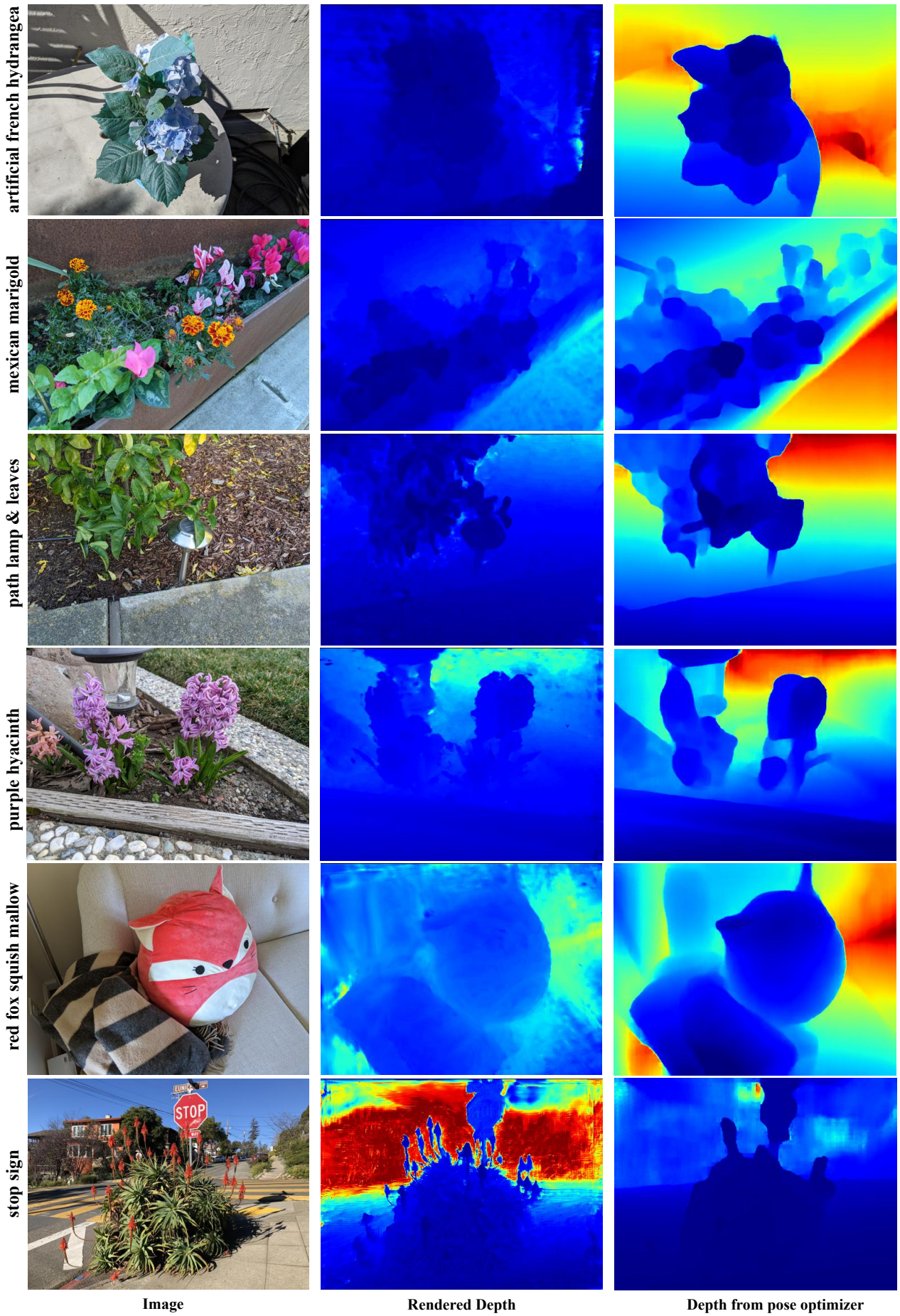

Figure 6. **Depth visualization on ibrnet self-collected dataset [6].**

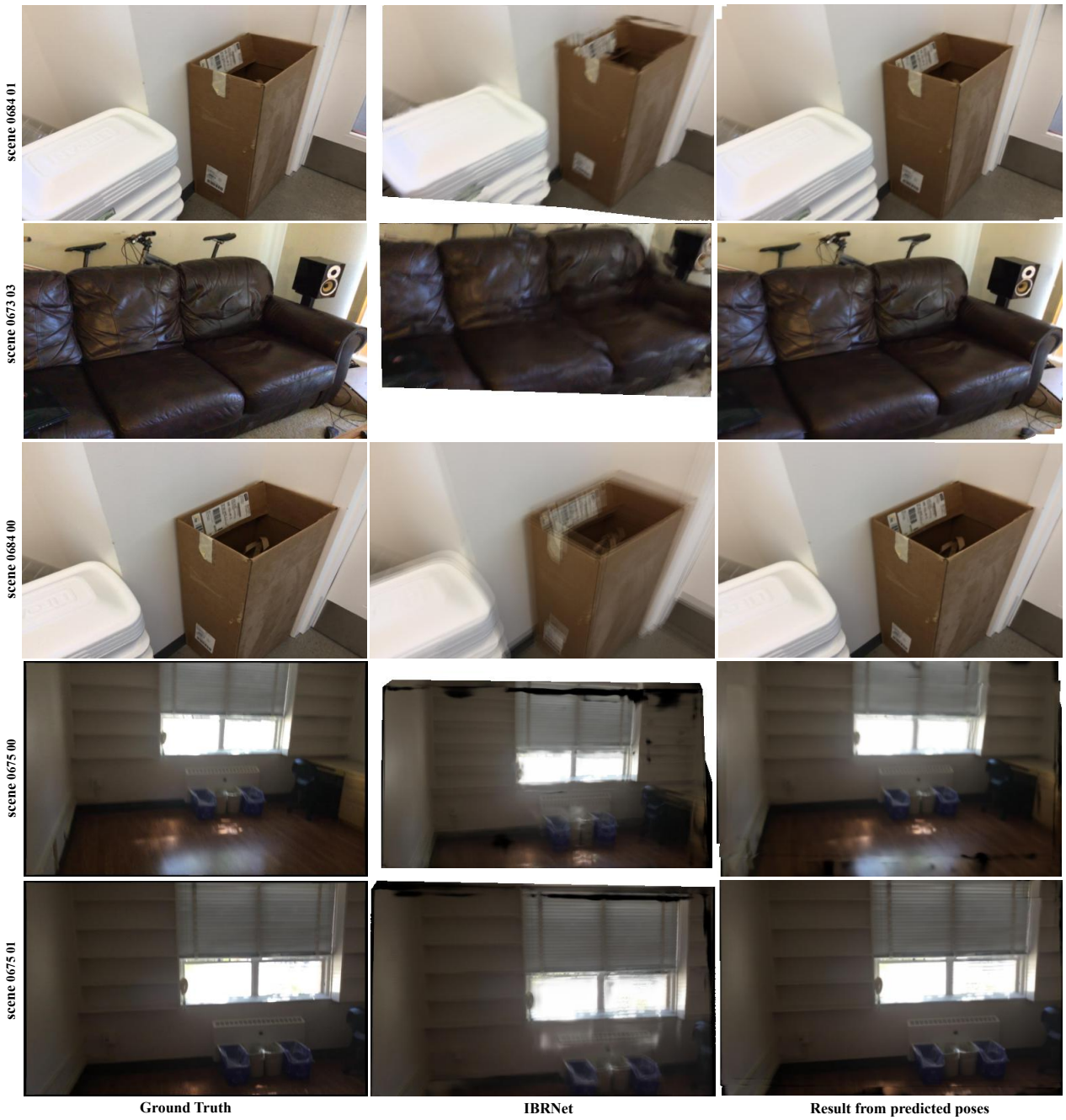

Figure 7. **The qualitative results on ScanNet dataset [2].**

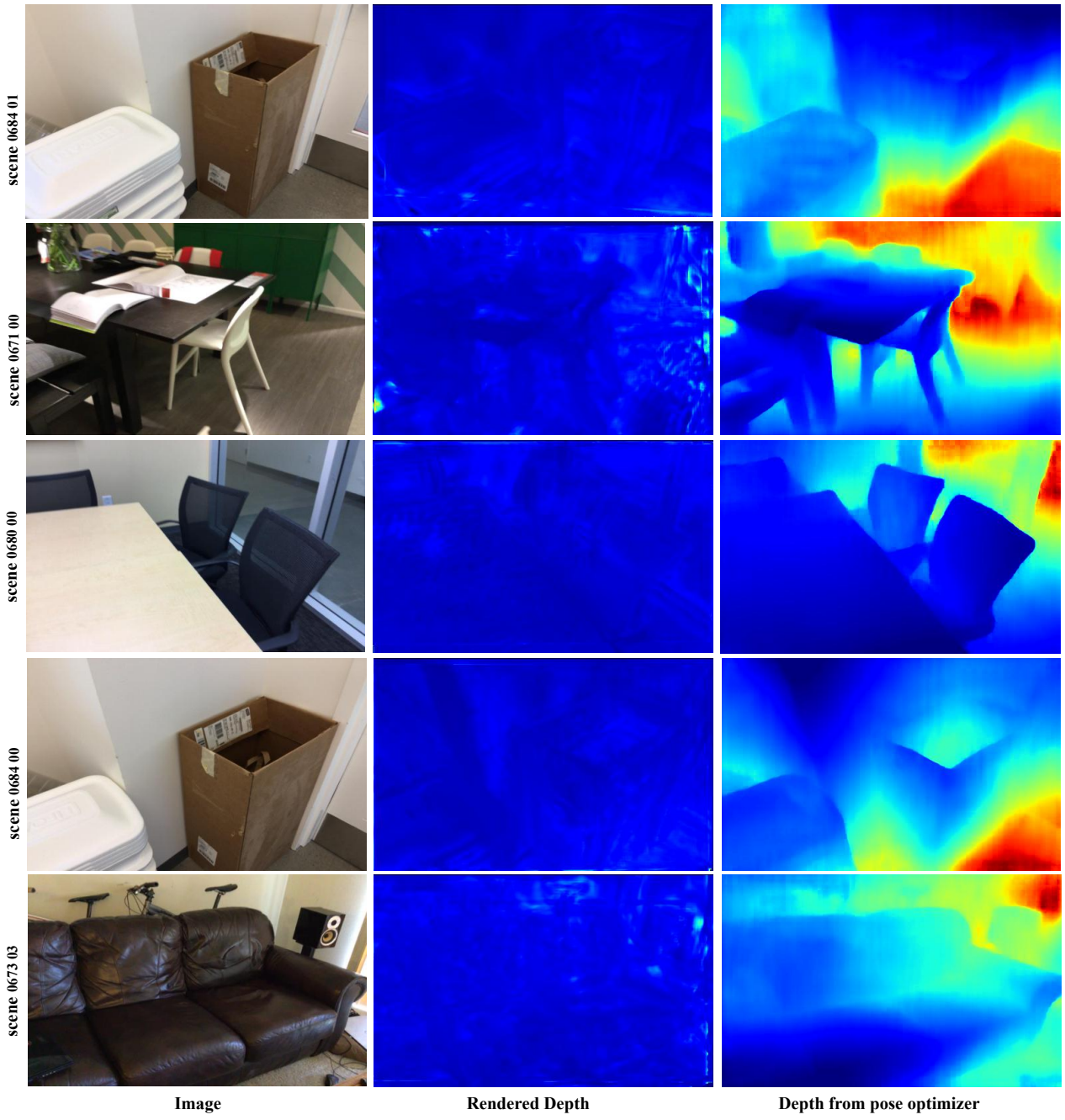

Figure 8. **Depth visualization on ScanNet dataset [2].**

References

[1] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. GARF: gaussian activated radiance fields for high fidelity reconstruction and pose estimation. *CoRR*, abs/2204.05735, 2022. 1

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443. IEEE Computer Society, 2017. 1, 6, 7

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE Computer Society, 2016. 1

[4] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: bundle-adjusting neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 5721–5731. IEEE, 2021. 1

[5] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Trans. Graph.*, 38(4):29:1–29:14, 2019. 1, 2

[6] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699. Computer Vision Foundation / IEEE, 2021. 1, 4, 5